Human Behavior Segmentation and Recognition Using a Single-camera

Jing Cao
Graduate School of Engineering, Kyushu Institute of Technology, 1-1 Sensui, Tobata-ku, Kitakyushu, 804-8550, Japan

Yui Tanjo
Faculty of Engineering, Kyushu Institute of Technology, 1-1 Sensui, Tobata-ku, Kitakyushu, 804-8550, Japan
Email: cao.jing644@mail.kyutech.jp, tanjo@cntl.kyutech.ac.jp

Abstract

In recent years, elderly people living alone account for a large proportion of the elderly population, and the issue of safety has also been a matter of great concern for the public. Considering the importance of monitoring the behavior and activities of the elderly and detecting abnormal movements, this paper proposes a method that can segment human behavior into each action and identify the action from the videos taken by a single camera. It uses features that can represent the shape of the human area in the depth direction, as well as the features such as motion direction and speed. The performance and effectiveness of the method are verified by experiments.

Keywords: Behavior segmentation, Motion recognition, Optical flows, TMRIs, Ex-HOOF, MHI

1. Introduction

Nowadays, the world's population is aging. The number and proportion of older people in the population is growing in every country in the world. Population aging will become one of the most significant social changes in the 21st century. By 2030, 1 in 6 people in the world will be 60 years or older (16%), and the share of people aged 60 and above will increase from 1 billion in 2020 to 1.4 billion. By 2050, the number of people aged 80 and over is expected to triple between 2020 and 2050, to 426 million [1]. Especially in Japan, 30% of the population is over 60 years old. Some social problems, such as solitary death which more than 50% of people worry about, are also intensifying [2]. Moreover, 67.1% of the respondents felt that the domestic security situation in Japan has deteriorated in the past 10 years [3]. Considering these problems, it is necessary to develop a system that can detect the behavior of the elderly and detect abnormal behaviors such as crime and theft, so as to realize a safe and secure society.

In related research on behavior recognition, X. Yang et al. [4] proposed a behavior segmentation and recognition method based on CSI, but it is not aimed at the segmentation of continuous human behavior. W. Xing et al. [5] proposed a behavior segmentation method based on posture histograms and adjusted sliding windows. This method requires learning the features of various postures. In related research on action recognition using computer vision, the conventional methods include the Flow Vectors method [6] and the MHI (Motion History Image) method [7]. However, these methods using a single camera are only suitable for the motion on the plane perpendicular to the optical axis of a camera, and the motion in the direction of the camera optical axis (toward or away from the camera) is not dealt with. Therefore, an extended 3D-MHI [8] and a reverse MHI method [9] have been proposed. However, [8] has the problem of high computational cost of creating 3D images, and, with [9], the recognition rate needs to be improved.

In this study, we propose a new method to describe the motion in the depth direction, making it possible to realize the segmentation and recognition of behaviors containing these motions. We segment human behavior by extracting features from the human area and then select key frames of the segmented actions. The key frames can be described through TMRIs (Triplet Motion Representation Images) [10]. The shape features of TMRIs, the features of optical flow and the changes of motion are also extracted to analyze and identify the motion of the frame. Finally, the recognition results of these frames are counted, and the final result of behavior recognition is obtained.

2. Methodology

2.1. *Behavior segmentation*

When human behavior consists of several types of motions, it is necessary to separate the behavior into respective motions to analyze the behavior. In this paper, we propose an automatic behavior segmentation method that does not require prior learning or training. The feature points are set at equal distances on the contour

line of the human area. Then feature points are tracked through the LK method [11] to obtain the optical flow, and RANSAC [12] is applied to find the true value by removing the influence of outliers. Changes in the direction of human movement can be distinguished from changes in the trajectory of the center of gravity coordinates and the average value of optical flow. Therefore, we calculate the point where a sudden change in trajectory occurs in the *x*-axis direction of the center of gravity coordinates and the boundary point between plus and minus of the average value of optical flow separately.

In Equation (1) and (2), if the change of the current frame exceeds a certain threshold, division is performed from the current frame and the number of the previous frame is used as the division point.

$$num_{tmp} = \begin{cases} num_{tmp} + 1 & if\ th1 < \dfrac{X\_dif_{sum}}{num_{tmp}} < th2 \\ 1 & otherwise \end{cases} \quad (1)$$

$$X\_dif_{sum} = \frac{\sum_{n=new}^{new-num_{tmp}} (x_n - x_{n-1})}{num_{tmp}} \quad (2)$$

where, *th1* and *th2* are thresholds. num_{tmp} is a parameter that determines whether or not to divide from the current frame. X_dif_{sum} is the average rate of change of the center of gravity's *x*-axis coordinate, and *new* is the number of the current frames.

When the direction of human motion changes, it can also respond to the average of optical flow. The intersection of the average optical flow m_{ave} and the *x*-axis is determined to be the division point. m_{ave} is defined by the following formula.

$$m_{ave} = \frac{\sum_{k=1}^{N} m_k}{N} \quad (3)$$

where, m_k represents the value of the *k*-th optical flow, and *N* denotes the total number of optical flows.

Then the division points are determined by matching the points of the two features after performing data smoothing within the threshold. The video can be divided using the obtained division points. After that, key frames are selected in each divided video, as shown in Fig. 1. In the figure, the orange points are the division points detected by features, which separate the behavior into several motions. The green points are the points that divide the motion into four equal parts, are also extracted as key frames.

In the end, by identifying key frames and determining the classification of the actions they represent through majority voting, the analysis of behavior composition is completed.

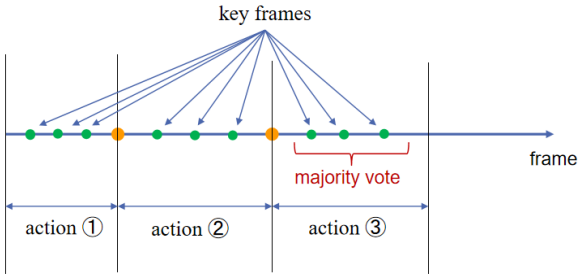


Fig.1 The definition of the ground truth.

2.2. *Motion recognition*

To capture the motion in the depth dimension, we employ TMRIs to represent actions. We identify and classify actions by leveraging their shape features, extracting directional velocities from the movements, and observing changes in the human region.

TMRIs is an extension of the conventional MHI. It is designed to express human motion, including motion along the camera optical axis, using three types of motion history images: *newness*, *density*, and *depth*. Among them, *newness* represents the MHI, *density* illustrates the appearance frequency of the foreground in the past τ frames, and *depth* indicates the depth information obtained from the FoE detection results. The shape features of TMRIs, is defined as $\boldsymbol{V}^{TMRIs}$. Ex-HOOF (Extended Histogram of Oriented Optical Flow) [10] is used to extract the speed and directional information of motions that are not readily visible in TMRIs features. The feature of Ex-HOOF is denoted as $\boldsymbol{V}^{Ex-HOOF}$.

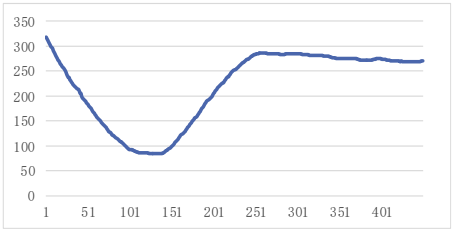
Additionally, to represent more complex and detailed motion features, we incorporate changes in the centroid and area (AC) of the human region. The feature can be defined as $\boldsymbol{V}^{AC} = (\boldsymbol{F}_A, \boldsymbol{F}_C)$. Among them, the area feature $\boldsymbol{F}_A = (Area_\tau^{ave}, Area_\tau^{sd})$, where $Area_\tau^{ave}$ is the average value, and $Area_\tau^{sd}$ denotes the standard deviation of the change of the foreground area. The formulas are as follows;

$$Area_\tau^{ave} = \frac{1}{\tau}\sum_{i=0}^{\tau} Area_{comp}^{t-i} \quad (4)$$

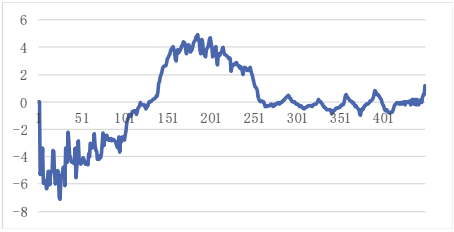
$$Area_\tau^{sd} = \sqrt{\frac{1}{\tau}\sum_{i=0}^{\tau} (Area_{comp}^{t-i} - Area_\tau^{ave})^2} \quad (5)$$

Here, $Area_{comp}^{t}$ is the changes of areas between the frames, given by

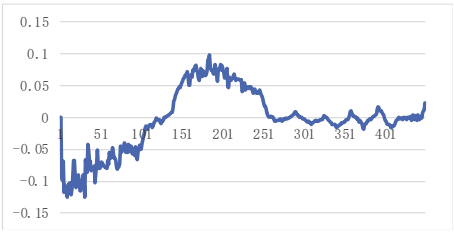
$$Area_{comp}^{t} = \frac{(Area_t - Area_{t-p})}{Area_t} \quad (6)$$



a. Change in the trajectory of the center of gravity coordinate in the *x*-axis



b. Average change in optical flow in the *x*-axis direction



c. Average change in optical flow in the *y*-axis direction

Fig.2. Features used for behavior segmentation.

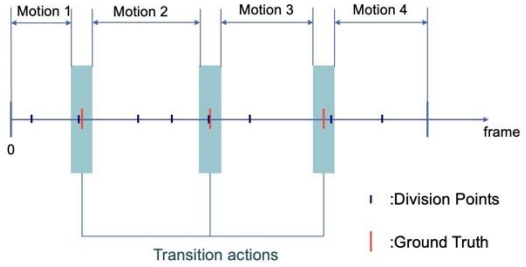


Fig.3 The definition of the ground truth.

The centroid feature $\boldsymbol{F}_C = (Cx_\tau^{ave}, Cy_\tau^{ave}, Cx_\tau^{sd}, Cy_\tau^{sd})$. C_τ^{ave} means the average value and C_τ^{sd} indicates the standard deviation of the change in the coordinate of the center of gravity. They are defined by

$$Cx_\tau^{ave} = \frac{1}{\tau}\sum_{i=0}^{\tau}(Cx_t - Cx_{(t-i-p)}) \tag{7}$$

$$Cy_\tau^{ave} = \frac{1}{\tau}\sum_{i=0}^{\tau}(Cy_t - Cy_{(t-i-p)}) \tag{8}$$

$$Cx_\tau^{sd} = \sqrt{\frac{1}{\tau}\sum_{i=0}^{\tau}((Cx_t - Cx_{(t-i-p)}) - Cx_\tau^{ave})^2} \tag{9}$$

$$Cy_\tau^{sd} = \sqrt{\frac{1}{\tau}\sum_{i=0}^{\tau}((Cy_t - Cy_{(t-i-p)}) - Cy_\tau^{ave})^2} \tag{10}$$

Finally, by integrating the above features, we get a feature vector $\boldsymbol{V}$, as shown below.

$$\boldsymbol{V} = (\boldsymbol{V}^{TMRIs}, \boldsymbol{V}^{Ex-HOOF}, \boldsymbol{V}^{AC}) \tag{11}$$

Action recognition is conducted using the *k*-nearest neighbors (*k*-NN) method in the proposed method.

3. Results and Discussion

3.1. *Accuracy of behavior segmentation*

In the behavior segmentation experiment, three people (labelled as 1,2,3) perform behavior A, B, C and D. The specific motion sequences are as follows: A: Walking left rear → Walking right → Walking front, B: Walking right rear → Walking left → Walking front, C: Walking rear → Walking left → Walking right front, D: Walking rear → Walking right → Walking left front. In the segmentation experiments, we verify the accuracy and recall of segmentation point selection based on the trajectory of the centroid in the *x*-axis and the average direction of optical flow. The example of the features of behavior A is shown in Fig. 2. The recall rate (*Recall*) is defined by

$$Recall = \frac{P_T}{P_{GT}} \times 100[\%] \tag{12}$$

Here, P_{ALL} is the total number of calculated segmentation points, P_T is the number of segmentation points that were accurately segmented, and P_{GT} is the number of ground truth segmentation points.

In the experiment, a motion between two motions, such as turning around or standing, is called a transition action. If the division point is within a transition action, this point is considered a correct division point. If one behavior consists of four consecutive motions, the definition of the ground truth division points is shown in Fig. 3.

Table 1 shows the results of behavior segmentation, yielding an average recall of 91.7%.

3.2. *Recognition rate of the motions*

In the motion recognition experiment, four people performed 12 kinds of motions, and features were created for these motions. In the experiment, we use leave-one-out cross-validation to evaluate the accuracy of motion recognition. The recognition rate *R* is defined by

$$R = \frac{N_T}{N_{ALL}} \times 100[\%] \tag{13}$$

Here, N_T is the total number of correctly recognized features and N_{ALL} represents the total number of features.

The recognition rates obtained using TMRIs, EX-HOOF, TMRIs + EX-HOOF, and the proposed method (TMRIs + EX-HOOF + AC) are shown in Table 2.

As shown in the table, the proposed method achieved an average recognition rate of 97.25%. This is because TMRIs represent depth information determined by FoE detection results. In addition, Ex-HOOF includes movement speed and direction information, and the detailed feature AC expresses changes in the area of the motion region and the center of gravity. Therefore, in the final proposed method, the average recognition rate for daily activities is 99.1%, and the average recognition rate for falling activities is 89.84%. Furthermore, the average recognition accuracy for motions that include movement in the depth direction is 96.60%, and the average recognition accuracy for other motions is 94.85%. The average recognition accuracy of the proposed method using TMRIs + EX-HOOF + AC is about 17.41% higher than the method using only TMRIs.

Table 1. The result of behavior segmentation

Person	Behavior	Number of division points				*Recall* [%]
		Trajectory	Optical flow	Integration	Positive number	
1	A	23	9	6	3	100.0
	B	33	20	9	4	100.0
	C	46	17	10	4	100.0
	D	27	17	7	4	100.0
2	A	21	11	5	3	75.0
	B	17	10	4	3	75.0
	C	28	9	5	4	100.0
	D	30	19	6	4	100.0
3	A	28	18	6	3	75.0
	B	20	16	6	4	100.0
	C	27	10	7	4	100.0
	D	28	14	6	3	75.0
Average						91.7

Table 2. The recognition rate of the motions

Motion	Recognition rate [%]			
	TMRIs	Ex-HOOF	TMRIs + Ex-HOOF	TMRIs + Ex-HOOF + AC
Walk left	87.5	44.4	93.1	99.4
Walk right	70.3	93.1	99.4	100.0
Walk front	78.1	75.3	99.4	99.7
Walk back	73.1	97.8	100.0	100.0
Walk left front	83.8	96.9	96.3	99.1
Walk right front	87.8	72.8	94.7	100.0
Walk left rear	83.4	92.2	96.6	98.8
Walk right rear	83.1	94.4	95.3	95.9
Fall left	91.9	63.8	91.9	91.9
Fall right	79.4	71.3	75.6	88.1
Fall front	61.9	44.4	80.6	89.4
Fall rear	69.4	63.1	86.3	90.0
Average	79.8	78.8	94.2	97.3

4. Conclusion

In this paper, we proposed a human behavior segmentation and recognition method, which can handle the behavior including movements in the depth direction. For each behavior, the segmentation method proposed in the paper effectively identifies division points, enabling the recognition of segmented actions. The average recall rate for behavior segmentation has reached 91.69%. Furthermore, in action recognition, the proposed method achieved an average recognition rate of 97.25%. This strongly validates the effectiveness of the proposed approach.

In the future, our focus will be on achieving a high-speed and fully automated process of behavior segmentation through recognition. Additionally, despite the current good recall rate, there is still a need to enhance the accuracy of the automatic segmentation method to reduce the subsequent workload of recognition and segmentation adjustments.

References

1. United Nations, Department of Economic and Social Affairs, Population Division, World Population Prospects 2022. (Online Edition).
2. Cabinet Office, 2022 White Paper on Aging Society (Overall version), 2022, pp.2-6. (in Japanese)
3. Cabinet Office, Outline of the 2022 Public Opinion Poll on Security, 2022, pp.24. (in Japanese)
4. X. Yang, J. Cheng, X. Tang et al., CSI-based human behavior segmentation and recognition using commodity Wi-Fi. J Wireless Com Network 2023(46), 2023.
5. W. Xing, W. Wang et al., A Novel Method for Automated Human Behavior Segmentation. Apr. 2016.
6. E.L. Andrade, R.B. Fisher, and S. Blunsden, Detection of emergency events in crowded scenes, Proceedings of IEEE International Symposium on Imaging for Crime Detection and Prevention, Hong Kong, China, 2006, pp.528-533.
7. A. Bobick, J. Davis, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.23(3), 2001, pp.257-267.
8. Y. Yamashita, J.K. Tan, and S. Ishikawa, Human motion description and recognition under arbitrary motion direction, Proceedings of SICE Annual Conference 2017, Kanazawa, Japan, 2017, pp.110-115.
9. J.K. Tan, S. Okae, Y. Yamashita, and Y. Ono, A method of describing a self-occlusive motion – A reverse motion history image, International Journal of Biomedical Soft Computing and Human Sciences, vol.24(1), 2019, pp.1-7.
10. J. Cao, Y. Yamashita, J.K. Tan, Human motion recognition using TMRIs with extended HOOF, Journal of Robotics, Networking and Artificial Life, vol.7(4), 2021, pp.231-235.
11. B.D. Lucas, T. Kanade, An Iterative Image Registration Technique with an Application to Stereo Vision, Proc. of Int. Joint Conf. on Artificial Intelligence, 1981, pp.674-679.
12. M. A. Fischer, R. C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, 1981, pp.381-395.

Authors Introduction

Ms. Jing Cao

She received B.E. from Republic of China and M.E. from the Graduate School of Engineering, Kyushu Institute of Technology, Japan in 2020. She is acquiring the D.E. in the same University. Her research interest includes computer vision, machine learning and motion recognition.

Dr. Yui Tanjo

She is currently a professor with the Department of Mechanical and Control Engineering, Kyushu Institute of Technology. Her current research interests include ego-motion analysis by MY VISION, three-dimensional shape/motion recovery, human detection, and its motion analysis from video. She was awarded SICE Kyushu Branch Young Author's Award in 1999, the AROB Young Author's Award in 2004, the Young Author's Award from IPSJ of Kyushu Branch in 2004, and the BMFSA Best Paper Award in 2008, 2010, 2013 and 2015. She is a member of IEEE, The Information Processing Society, The Institute of Electronics, Information and Communication Engineers of Japan.